

**Introduction to Data Management PROJECT REPORT**

(Project Semester August-December 2021)

**PROJECT REPORT ON**

## Goodreadsbooks data

Submitted by

**BACHU VARUN**

**11912352**

Program: Bachelor of Technology Section: KM006

Course Code: INT217

Under the Guidance of **Komal Arora: 17783 Assistant Professor**

**Discipline of CSE/IT**

**Lovely School of Computer Science & Engineering Lovely Professional University, Phagwara**

# DECLARATION

I, Bachu Varun, student of Computer Science & Engineering under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

**Date: 18/12/2021**

**Bachu Varun**

**Registration No: 11912352**

**Signature: Whiteboard

Description automatically generated with low confidence**

# ACKNOWLEDGEMENT

Primarily I'd thank God for being able to complete my project with success. Then I'd like to thank my mentor **Ms. Komal Arora**, whose valuable guidance has been the ones that helped me patch this project and make it full proof success in contribution towards the completion of this project.

Finally, I'd rather thanks to **Lovely Professional University,** and my parent’s inspiration, who gave me this golden opportunity to learn many new things, to learn another aspect of life.

**-Bachu Varun**

# CONTENTS:

|  |  |  |
| --- | --- | --- |
| ***Sr No*.** | ***Title*** | ***Page No.*** |
| *1* | *Introduction* | *5* |
| *2* | *Objectives/Scope of the Analysis* | *6* |
| *3* | *Source of dataset* | *7-8* |
| *4* | *ETL Process* | *9-21* |
| *5* | *Analysis of dataset* | *22-31* |
| *6* | *List of Analysis with results* | *32-36* |
| *7* | *Final Dashboard* | *37* |
| *8* | *Bibliography* | *38* |

**INTRODUCTION**

Goodreadsbooks is a data of some good books. It is very useful to the book-lovers out there, irrespective of language and publication and all of that. We can use it to our liking, find what book we should be reading next. This data consists of bookID, Title, Authors, Average\_rating, isbn13(International Standard Book Number), language\_code, number of pages, ratings\_count, text\_reviews\_count, publication\_date, publisher.

Tableau is a data analysis and visualization tool which is commonly used in today's industry. Many organizations still find it important for the research relevant to data science. The ease ofuse of Tableau is due to it providing a drag and drop interface. This feature helps to perform tasks like sorting, comparing, and analyzing, very easily and fast. Tableau is also compatible with multiple sources, including Excel, SQL Server, and cloud-based data repositories which makes it an excellent choice for Data Scientists.

# OBJECTIVES/SCOPE OF ANALYSIS

After analysis of the dataset, the aim of this project is to give answer of given objectives in easy way:

* *Total books published in each year*
* *Top 50 average rated books*
* *Most rating counts containing books in each year*
* *No. of books written in each language*
* *Top 10 text reviewed books from whole dataset*

# SOURCE OF DATASET:

**Source of dataset:**

https://www.kaggle.com/jealousleopard/goodreadsbooks

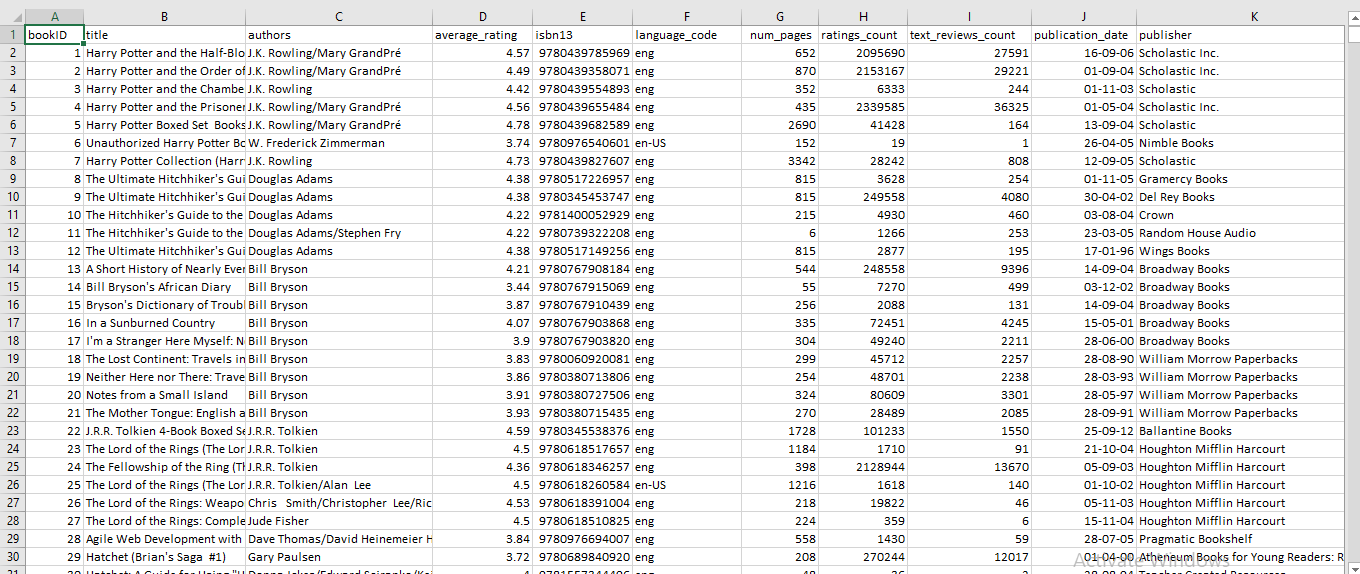
Kaggle is an online community for data scientists and machine learners, developed by Google. Kaggle allows users to find and publish data sets, explore, and build models in a web- based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges. Kaggle got its start by offering machine learning competitions and now also offers a public data platform, a cloud- based workbench for data science, and short form AI education. On 8 March 2017, Google announced that they were acquiring Kaggle.

This data science project analyzes the Goodreadsbooks dataset. It was created for the (B. Tech CSE fifth semester Introduction to Data Science course) project.

Every part of the dataset consists of multiple of punctuation errors which is cleaned in the ETL process.

### Sample of dataset with data fields is given below:

Here we can see multiple columns with names such as BookID, Title, Authors, Average rating, isbn13, language code, number of pages, text reviews count, publication date, publisher…so on



# ETL PROCESS:

ETL is defined as a process that extracts the data from different RDBMS source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the datainto the Data Warehouse system. ETL full form is Extract, Transform and Load.

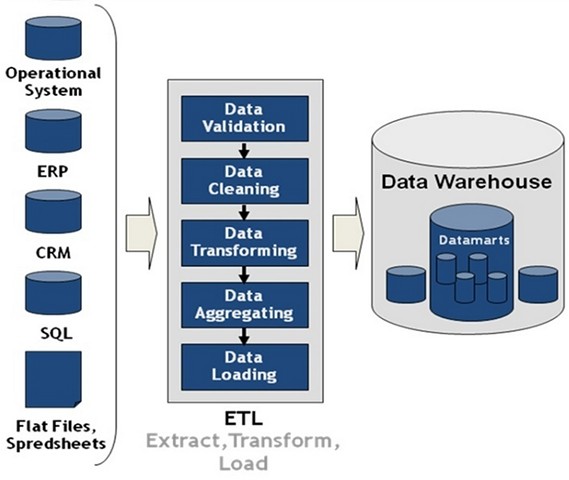
It's tempting to think a creating a Data warehouse is simply extracting data from multiple sources and loading into database of a Data warehouse. This is far from the truth and requires a complex ETL process. The ETL process requires active inputs from various stakeholders including developers, analysts, testers, top executives and is technically challenging.

## Need of ETL Process

* *ETL process allows sample data comparison between the source and the target system.*
* *ETL is a predefined process for accessing and manipulating source data into the targetdatabase.*
* *Allow verification of data transformation, aggregation, and calculations rules*.

When it comes to the implementation of the ETL process, the itinerary of tasks can be divvied up into the full form of its acronym.

1. **E – Extraction**
2. **T – Transformation**
3. **L – Loading**

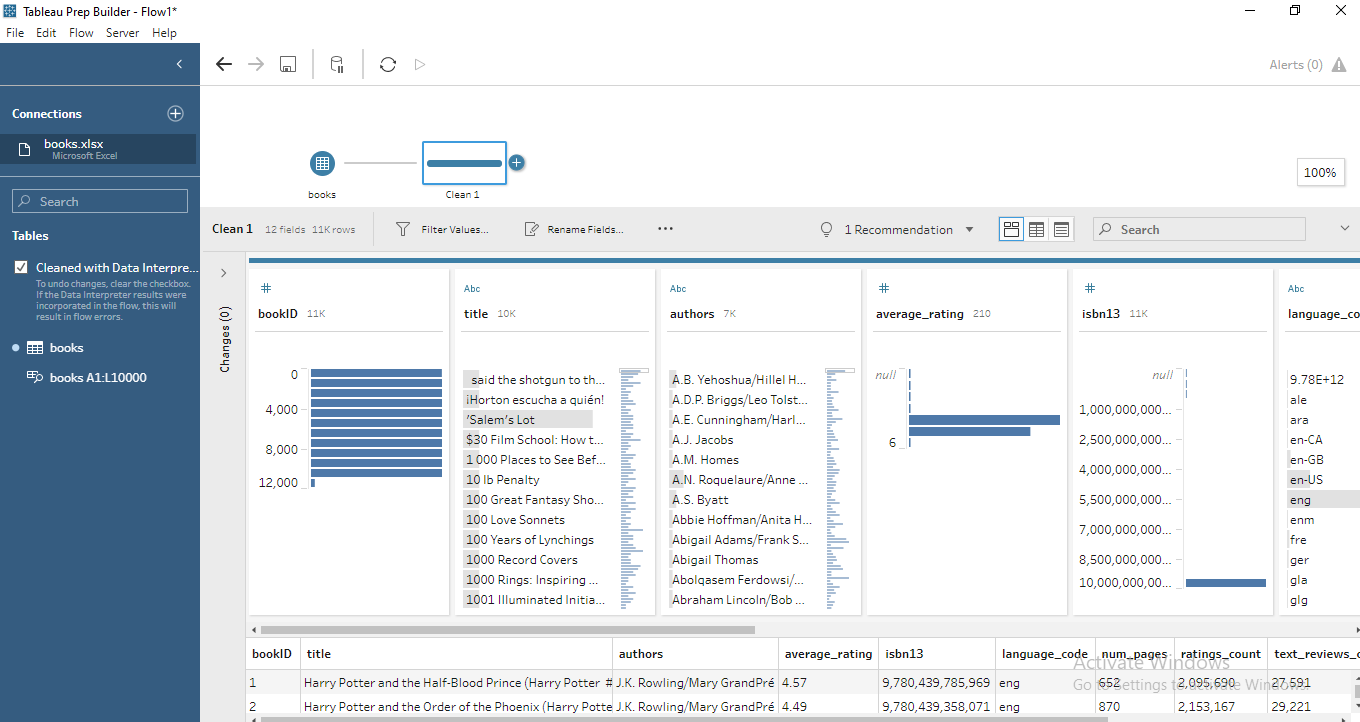


## ETL Process used in Project

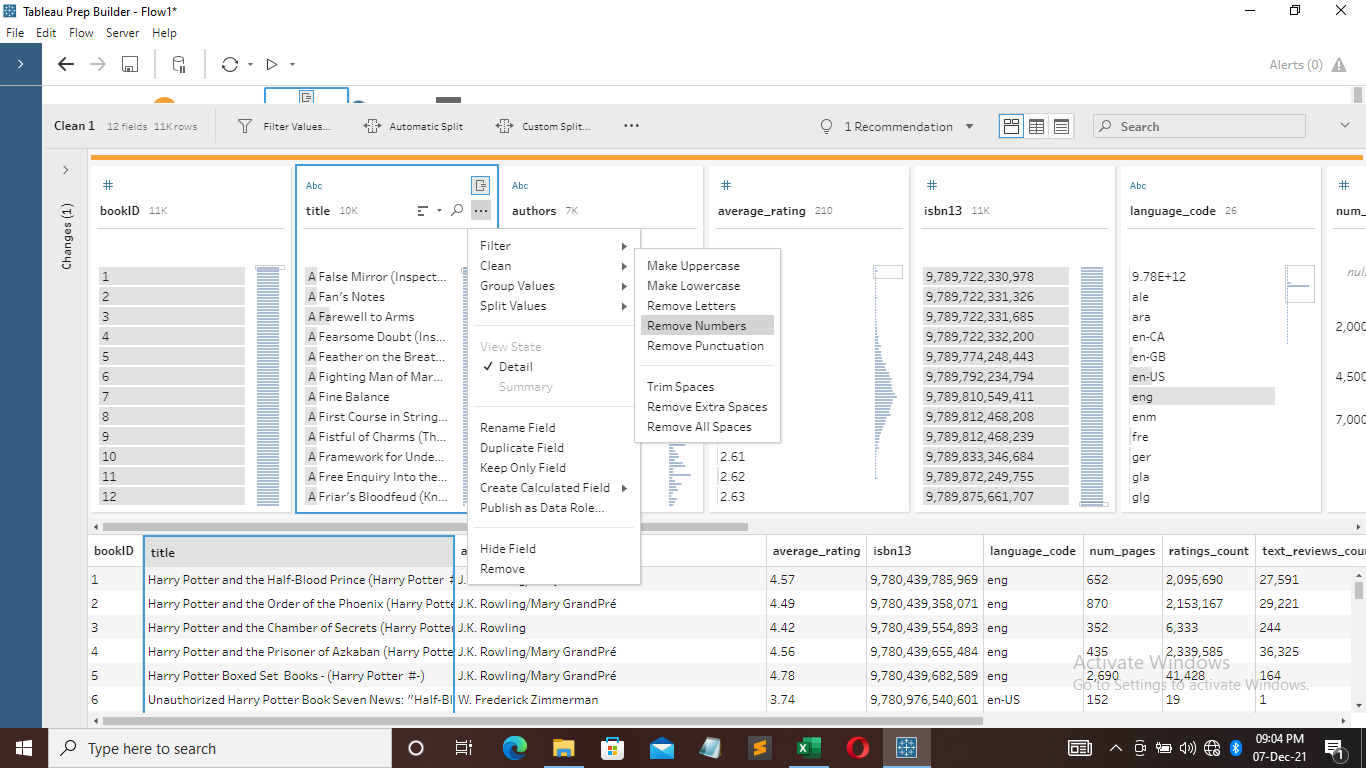
**Extraction**

Extracting the dataset from PC to Tableau for removing the unwanted characters, fields,spelling errors etc.

**Step 1:** *Opening the dataset in single table****.***



**Step 2:** Removal of numbers in Title column



**Step 3**: Removal of punctuations in title column

Graphical user interface, table

Description automatically generated

**Step 4**: Trimming spaces in title column

Graphical user interface, table

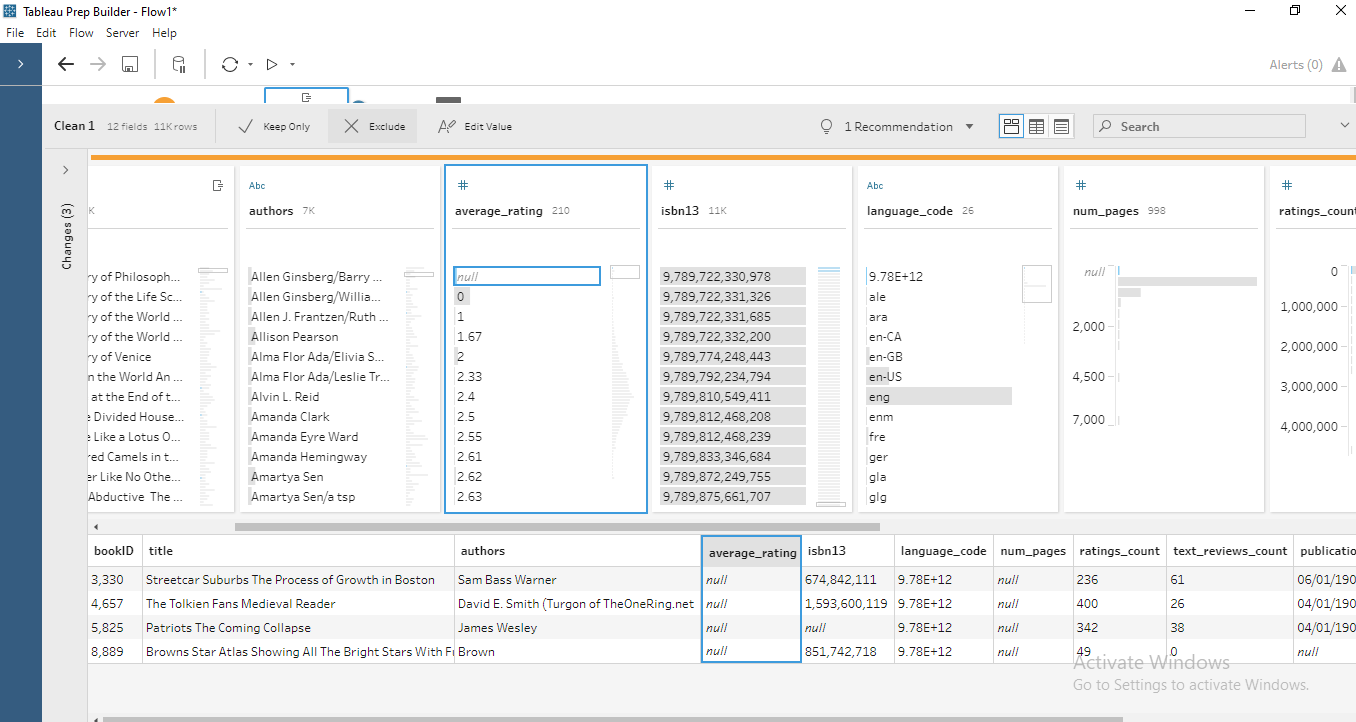
Description automatically generated

**Step 5:**  Removal of null values in average\_rating column

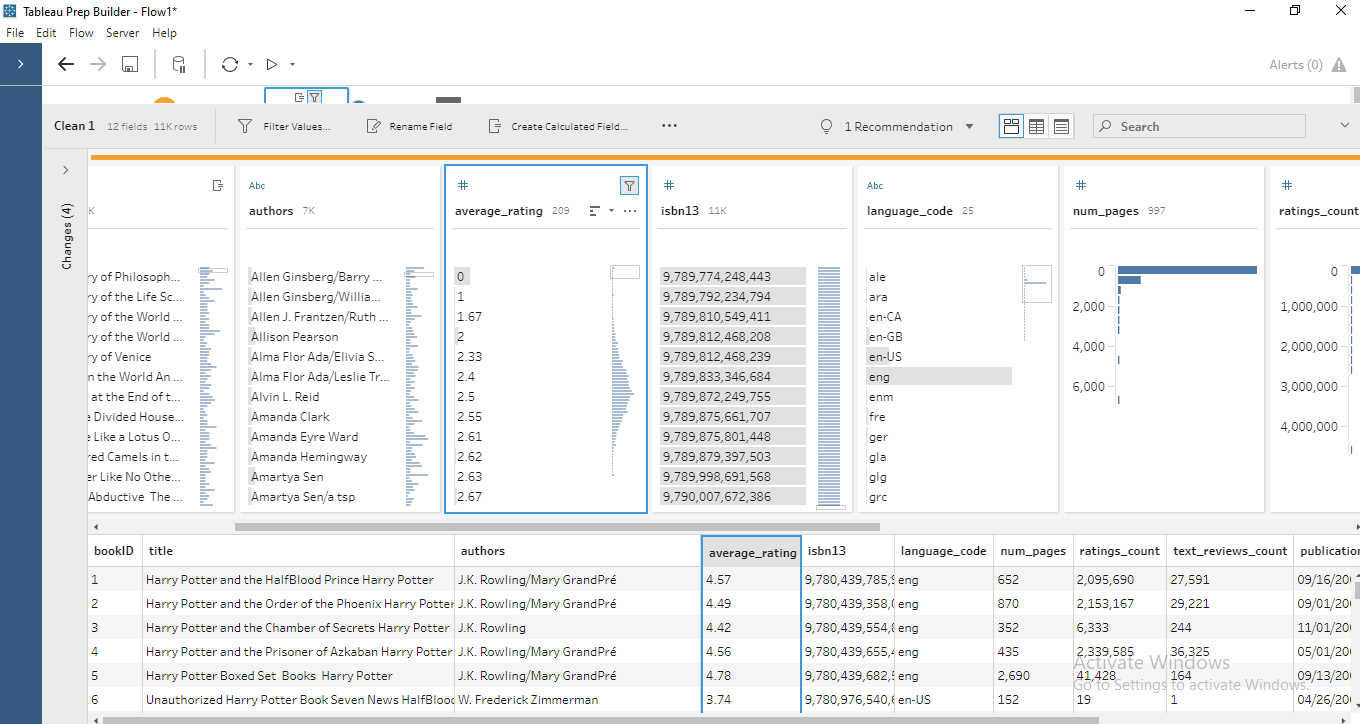
Graphical user interface, application, table

Description automatically generated

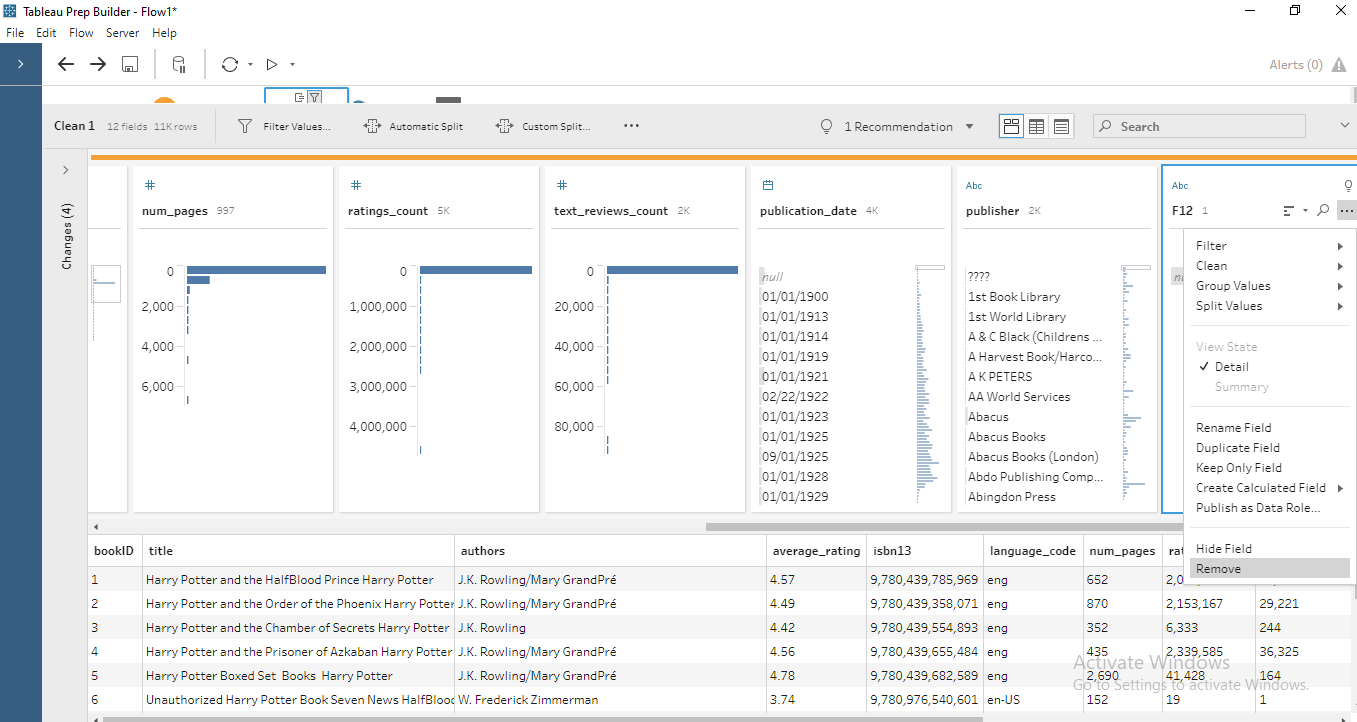
Selecting null values by manual selection and excluding it



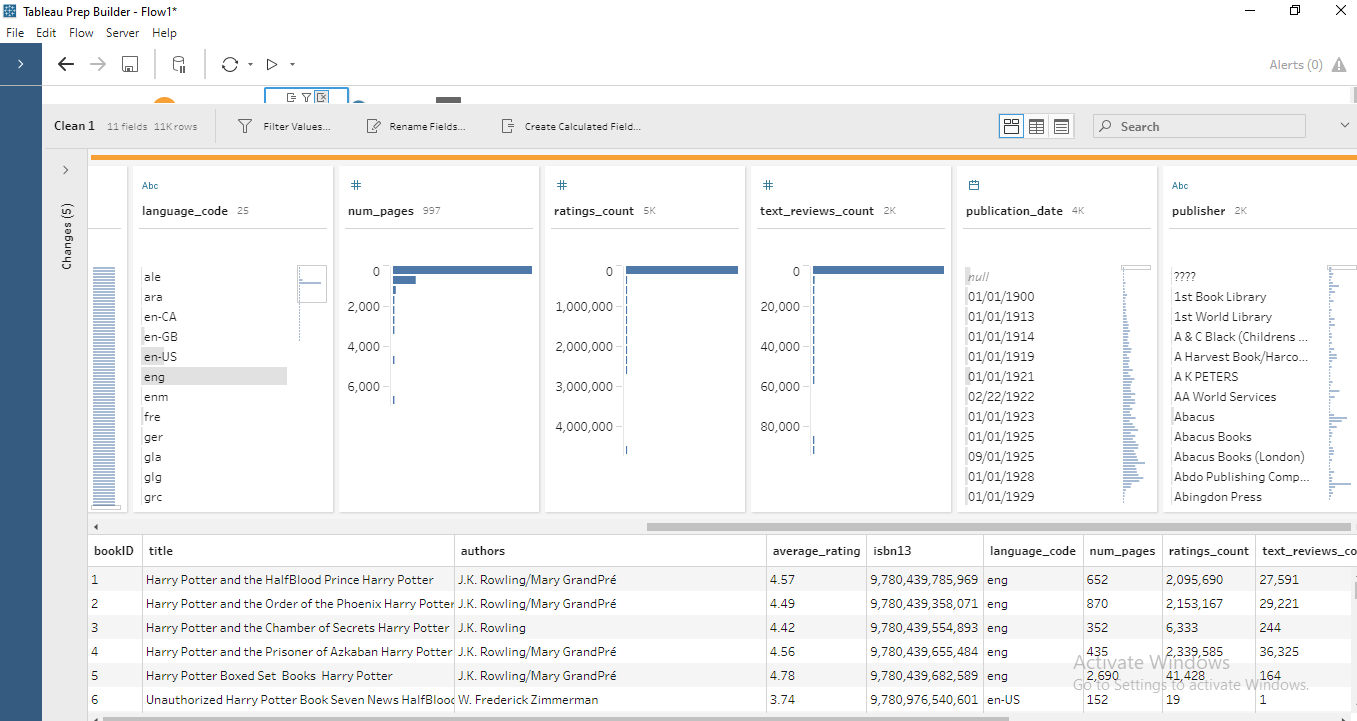
After removing null values



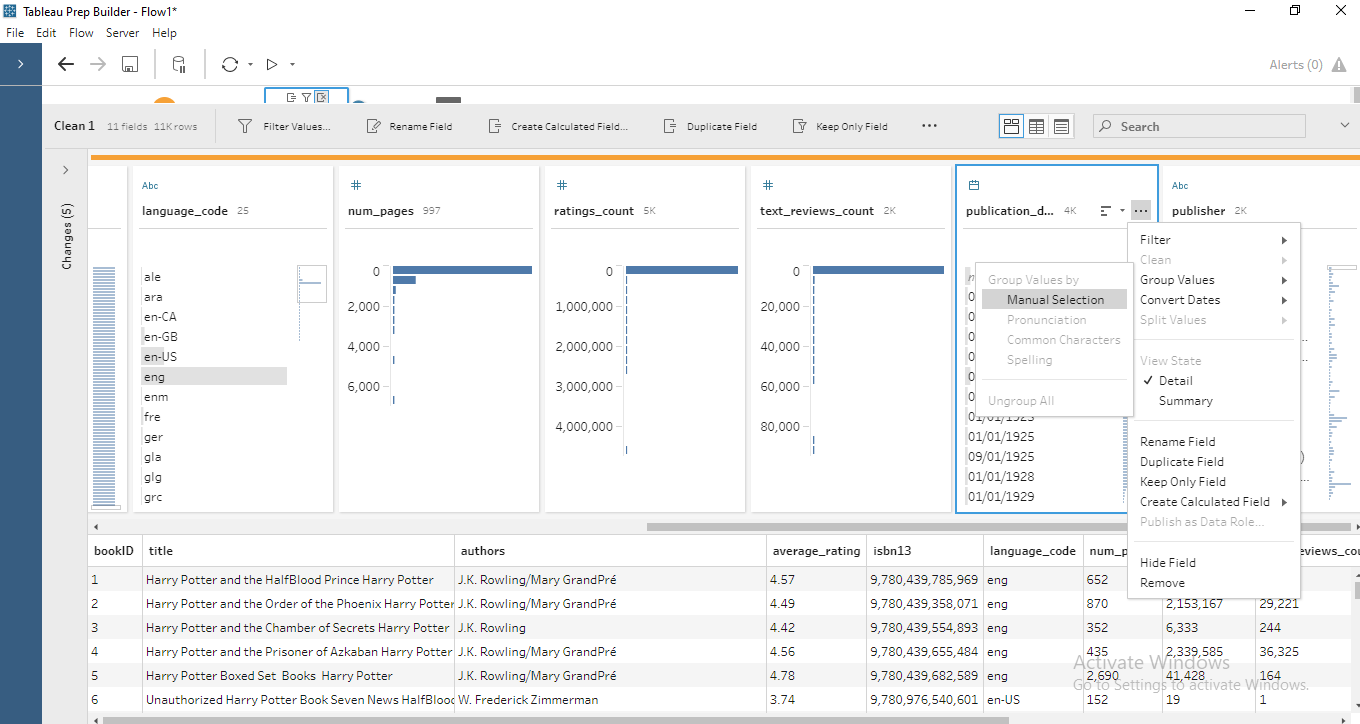
**Step 6**: Removal of f12(null) column



*After removing column:*

**

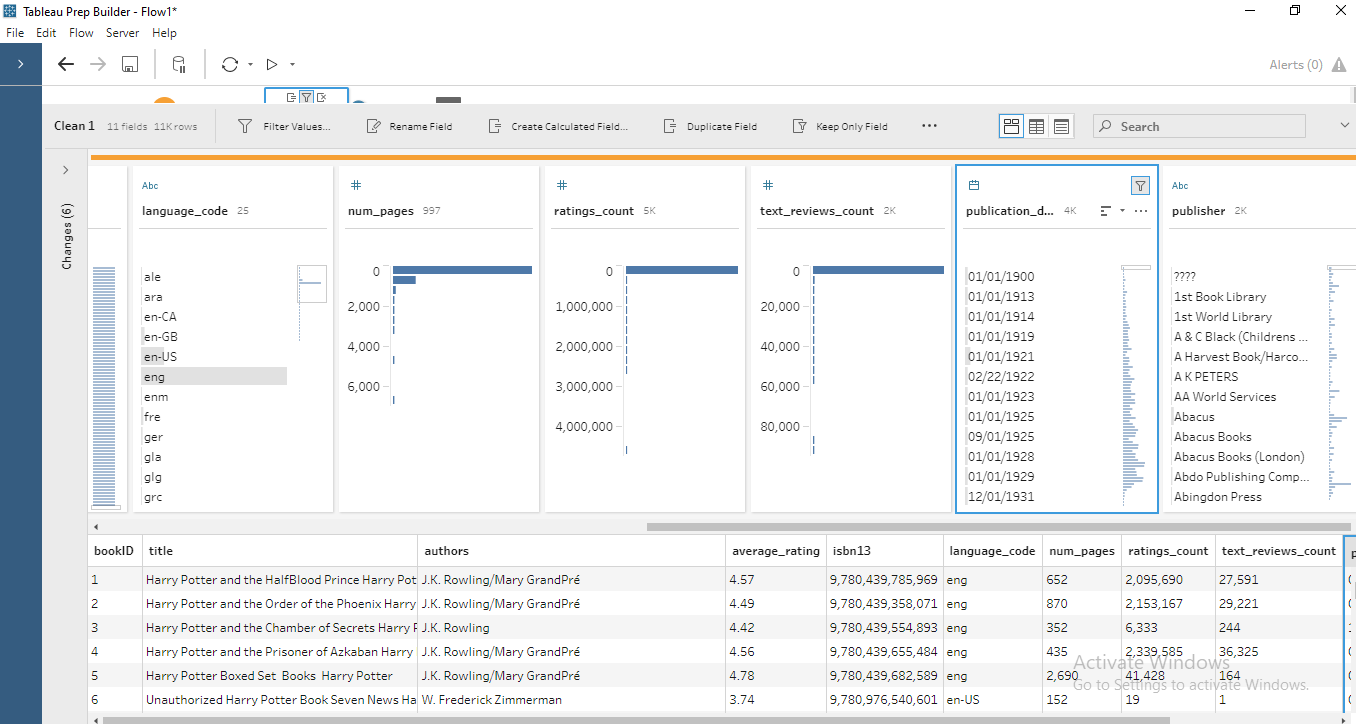
**Step-7**: *Removal of null values from publication\_date*



Selecting null values by manual selection and excluding it

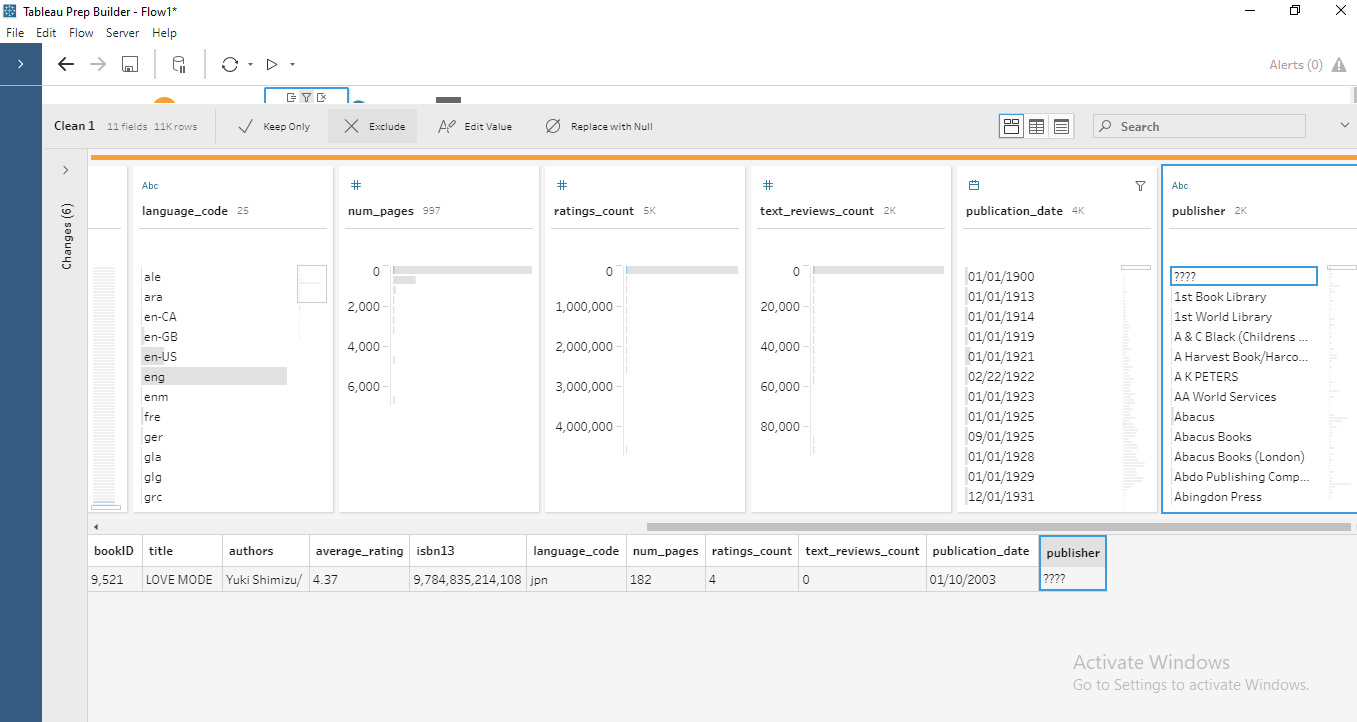


After removing null values:

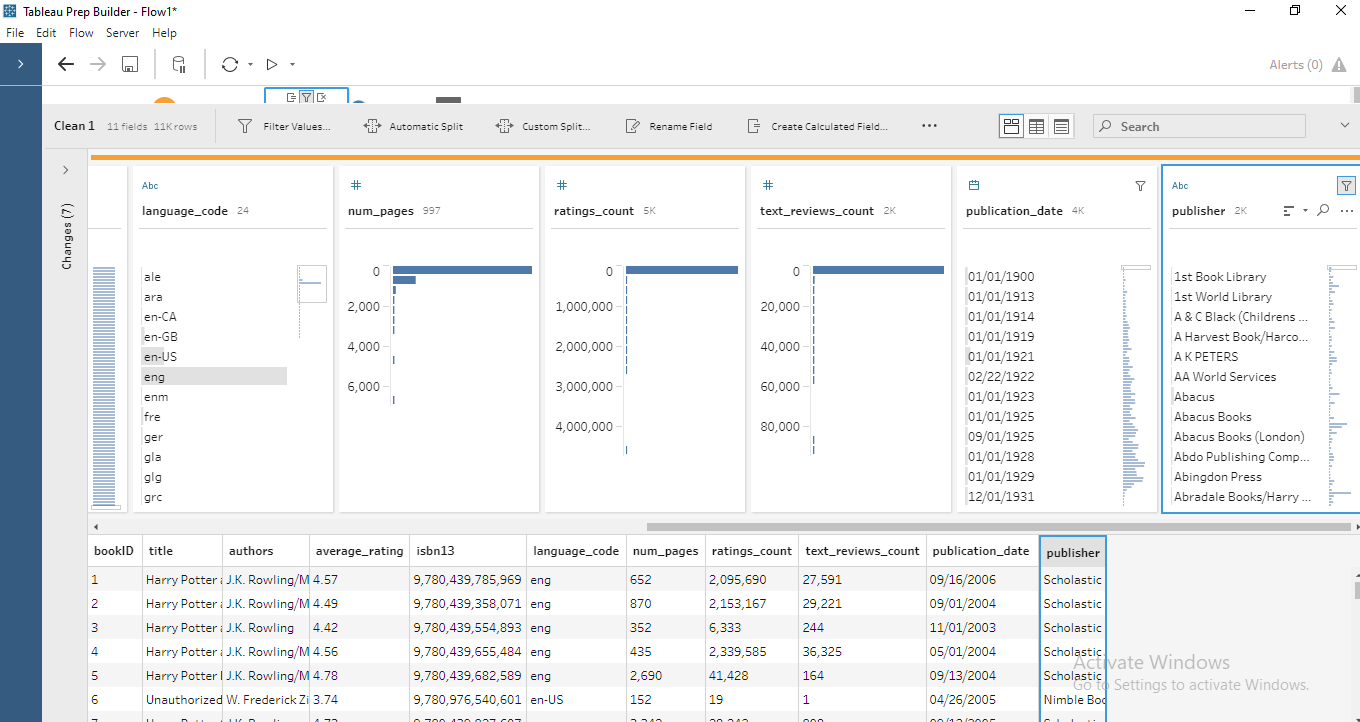


**Step-7**: Removing special characters from publisher

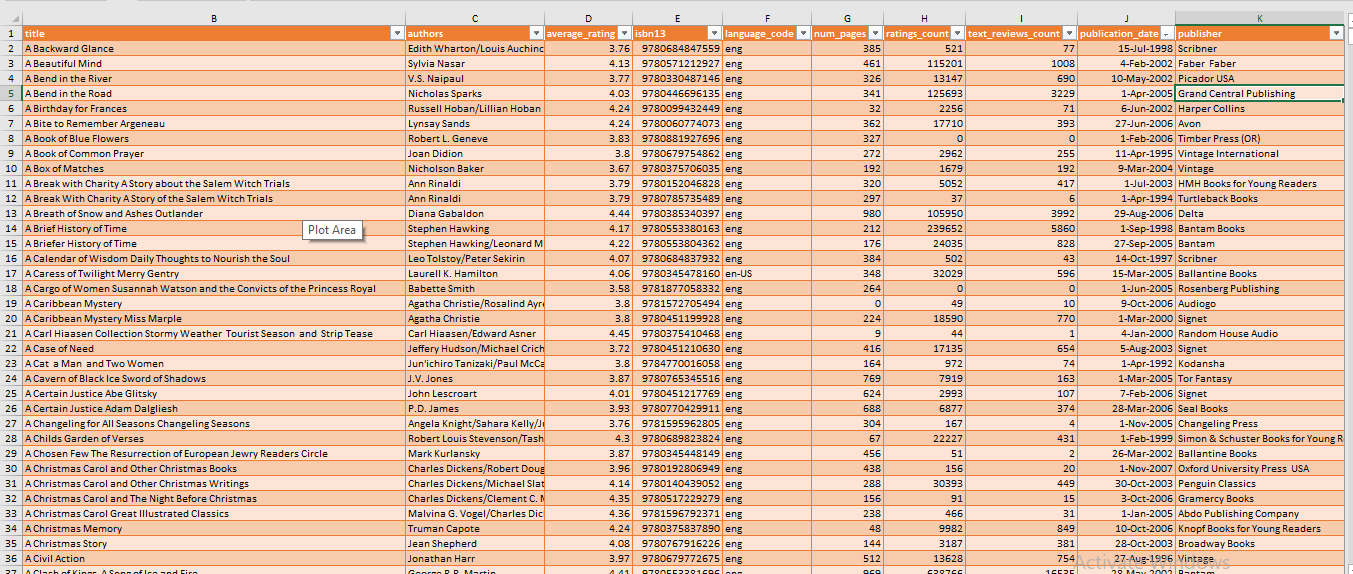
Selecting by manual and exclude:



After removing:



### Finally, after cleaning the data, the final dataset sample is shown below:

****

**Analysis on dataset**

1. ***Total books published in each year Introduction***
   * *By performing this analysis, we will get number of books published in each year*

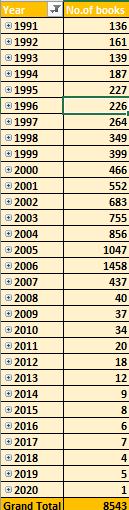
### Description:

It is to see the number of books published in each year.

### Specific requirements, functions, and formulas:

For counting number of books we can use count function: =count()

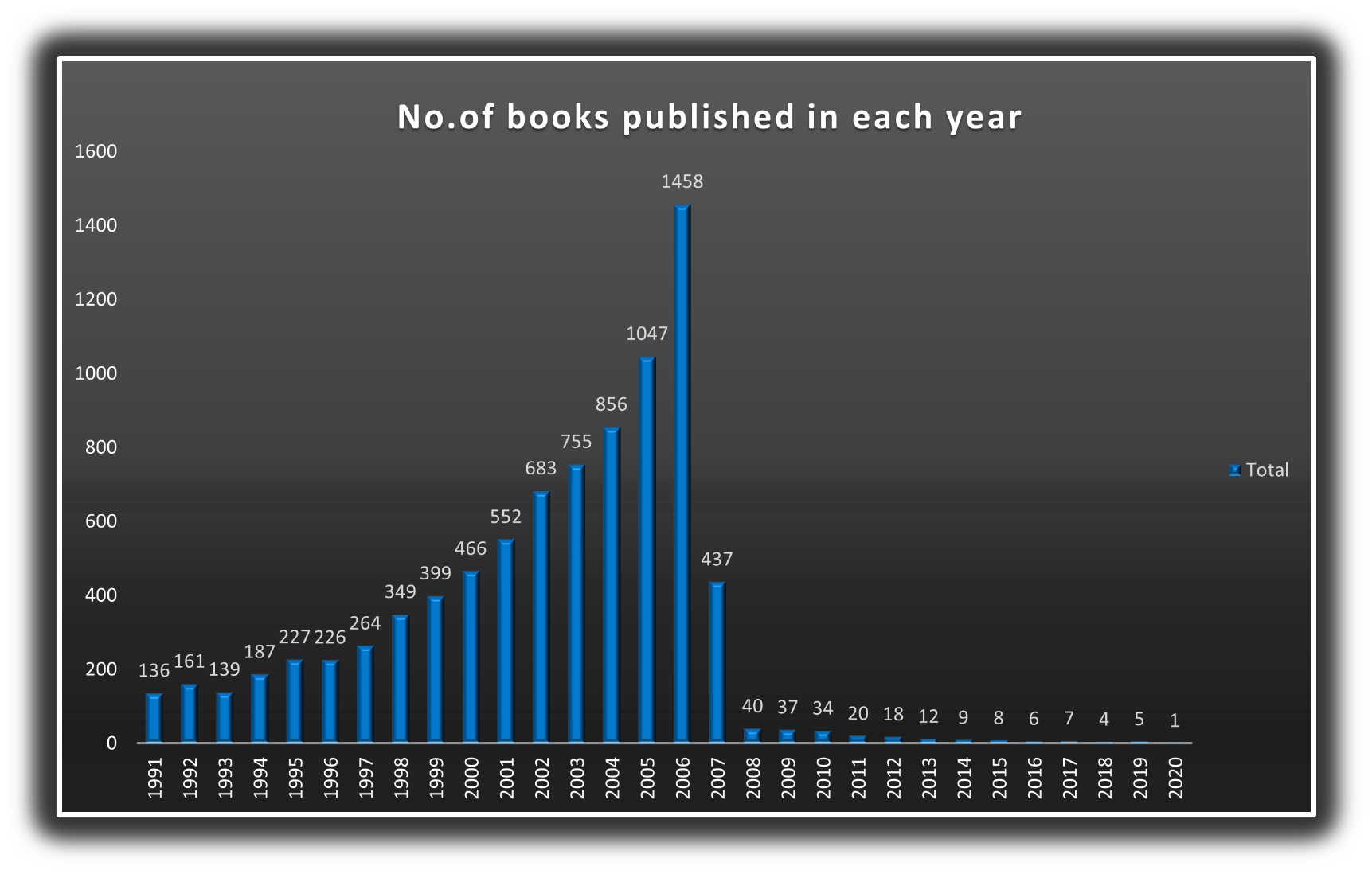
**Analysis results:** In the year 2006, most number of books published i.e., 1458.



### Slicer:

### 

***Visualization:***



1. ***Top 50 average rated books***

***Introduction***

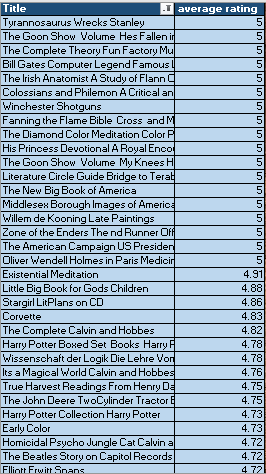
* + *By performing this analysis, we will get the top 50 books with respect to average rating*

### Description:

The analysis is based on top 50 rated books in the dataset with respect to average\_rating

### Analysis results:

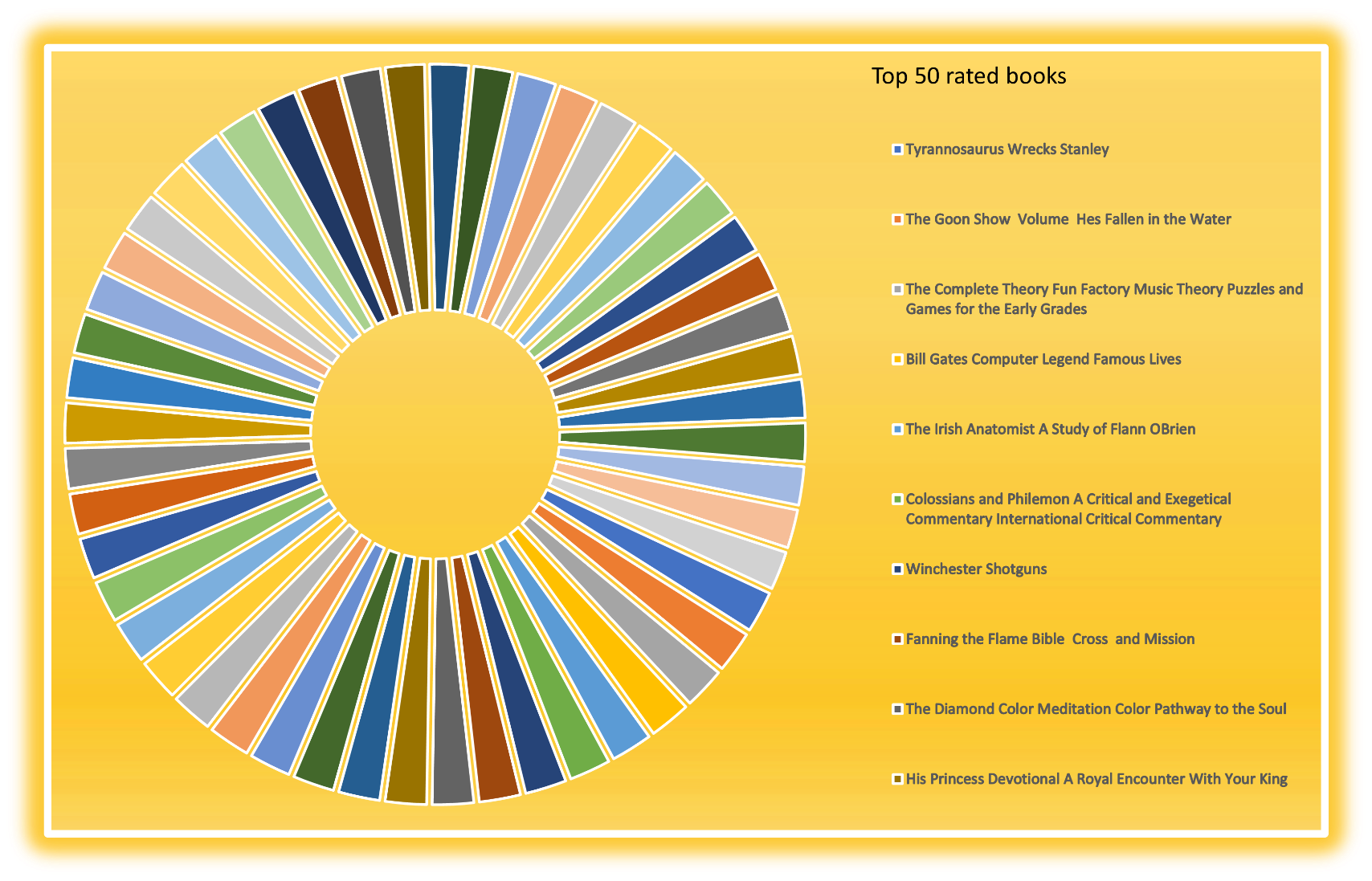
This analysis shows the top 50 books with top rating and names of books.



### Slicer:

### 

***Visualization:***

****

1. ***Most rating counts containing books in each year Introduction***

* *By performing this analysis, we will get maximum rating counts book in a particular year*

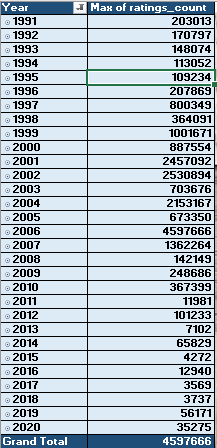
### Description:

The analysis based on maximum of ratings count containing books in each year

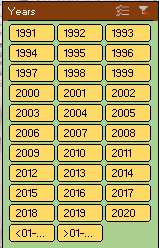
### Specific requirements, functions, and formulas:

Select the cell below the given Quantity and apply the formula ‘=max ()’. This function will give the maximum number in a range of cells

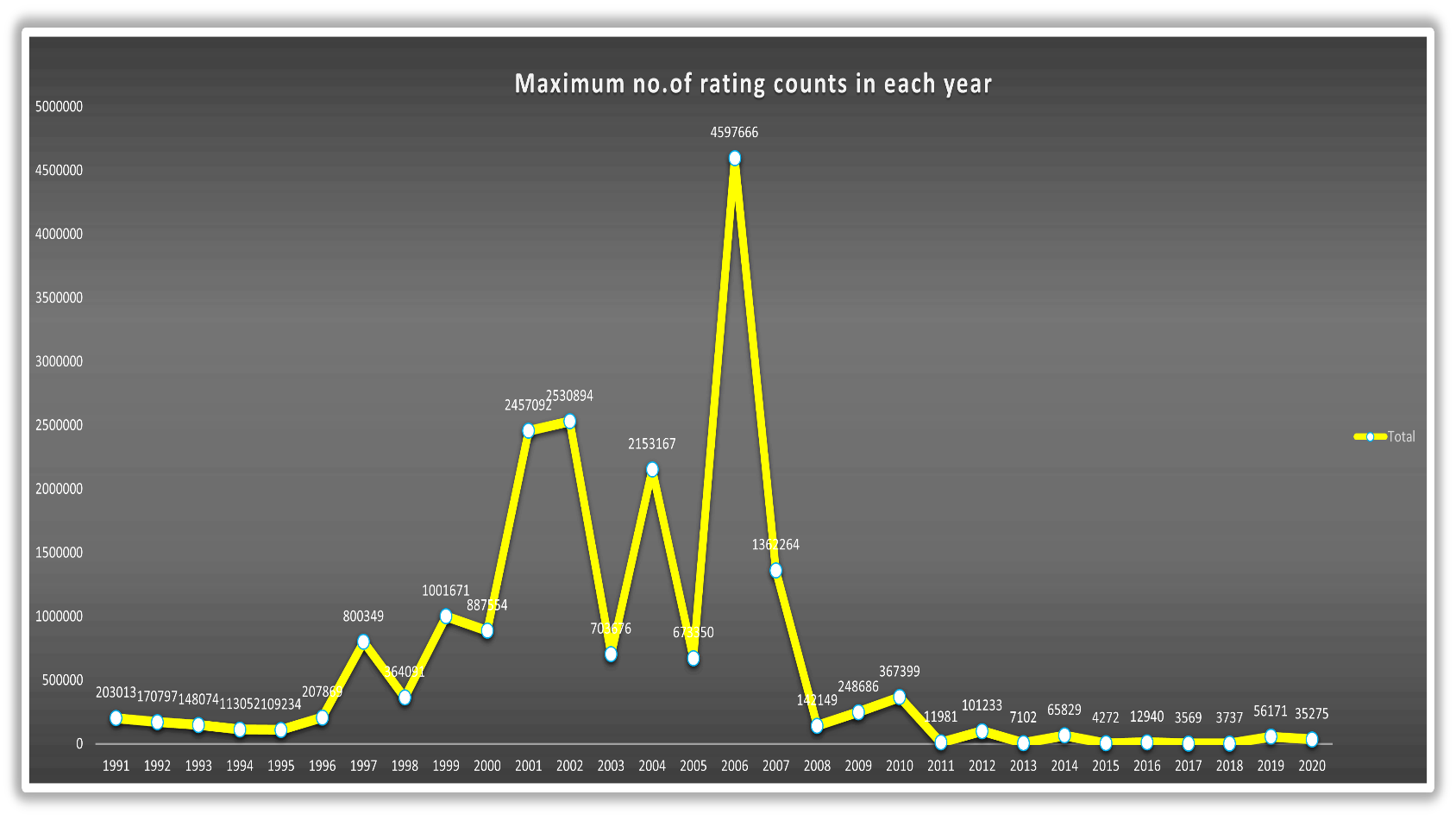
### Analysis results:

****

***Slicer:***

****

***Visualization:***

****

1. ***Number of books written in each language***

***Introduction***

* *By performing this analysis, we will get number of books written in each language with respect to language\_code*

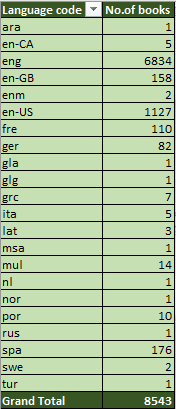
### Description:

### In this analysis, we display the language code and number of books written in that specific language.

### Specific requirements, functions, and formulas:

Select the cell below the given Quantity and apply the formula ‘=count()’. This function will count the number of books in a range of cells

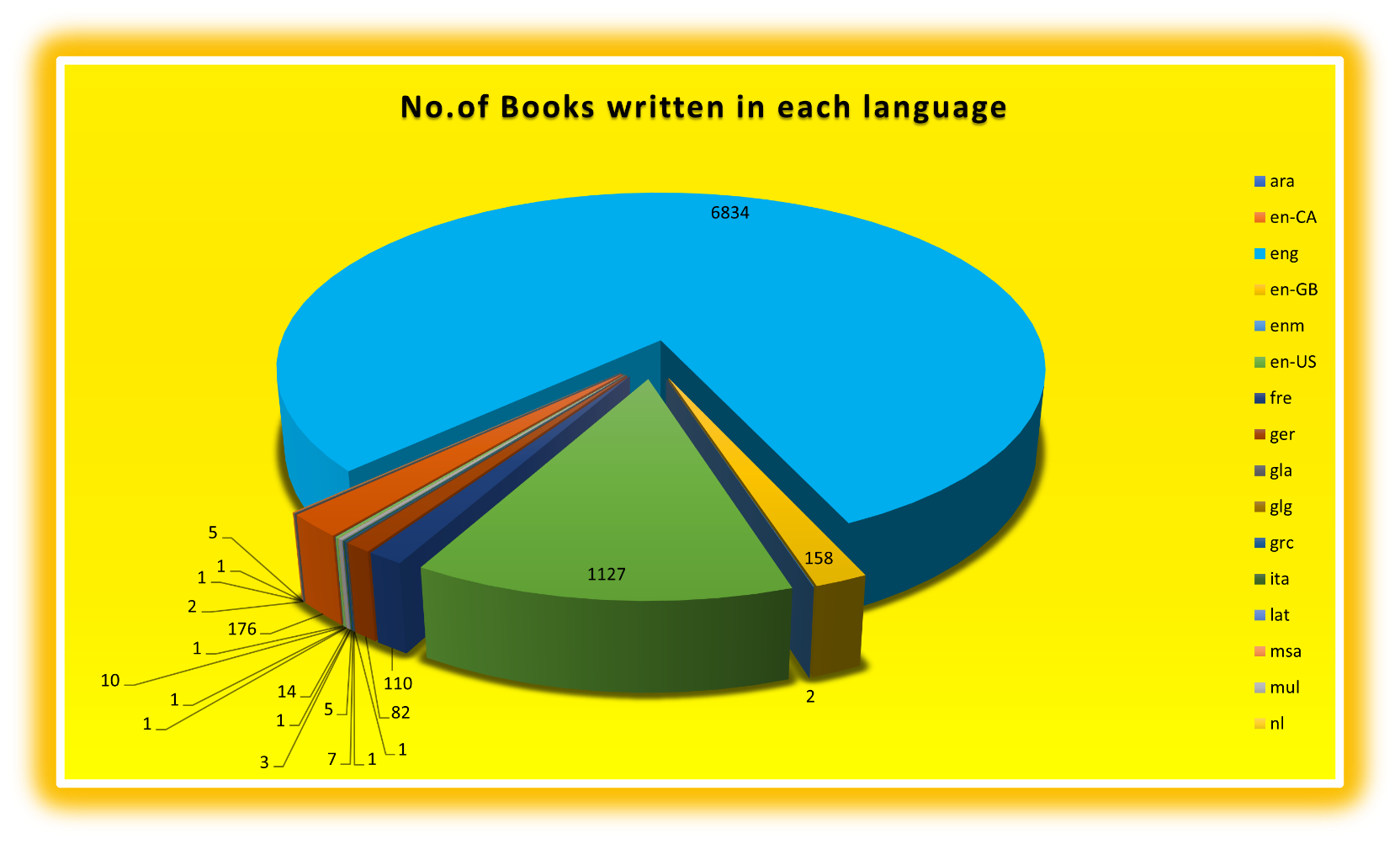
### Analysis results:



***Slicer:***

******

***Visualization:***



1. ***Top 10 text reviewed books***

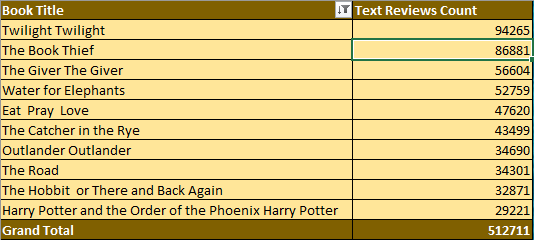
***Introduction***

* *By performing this analysis, we will get top 10 text reviewed books*

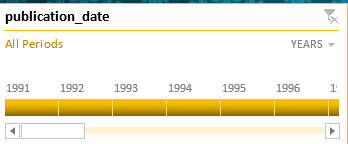
### Description:

10 Most text reviewed books gives which are the best books in the whole dataset.

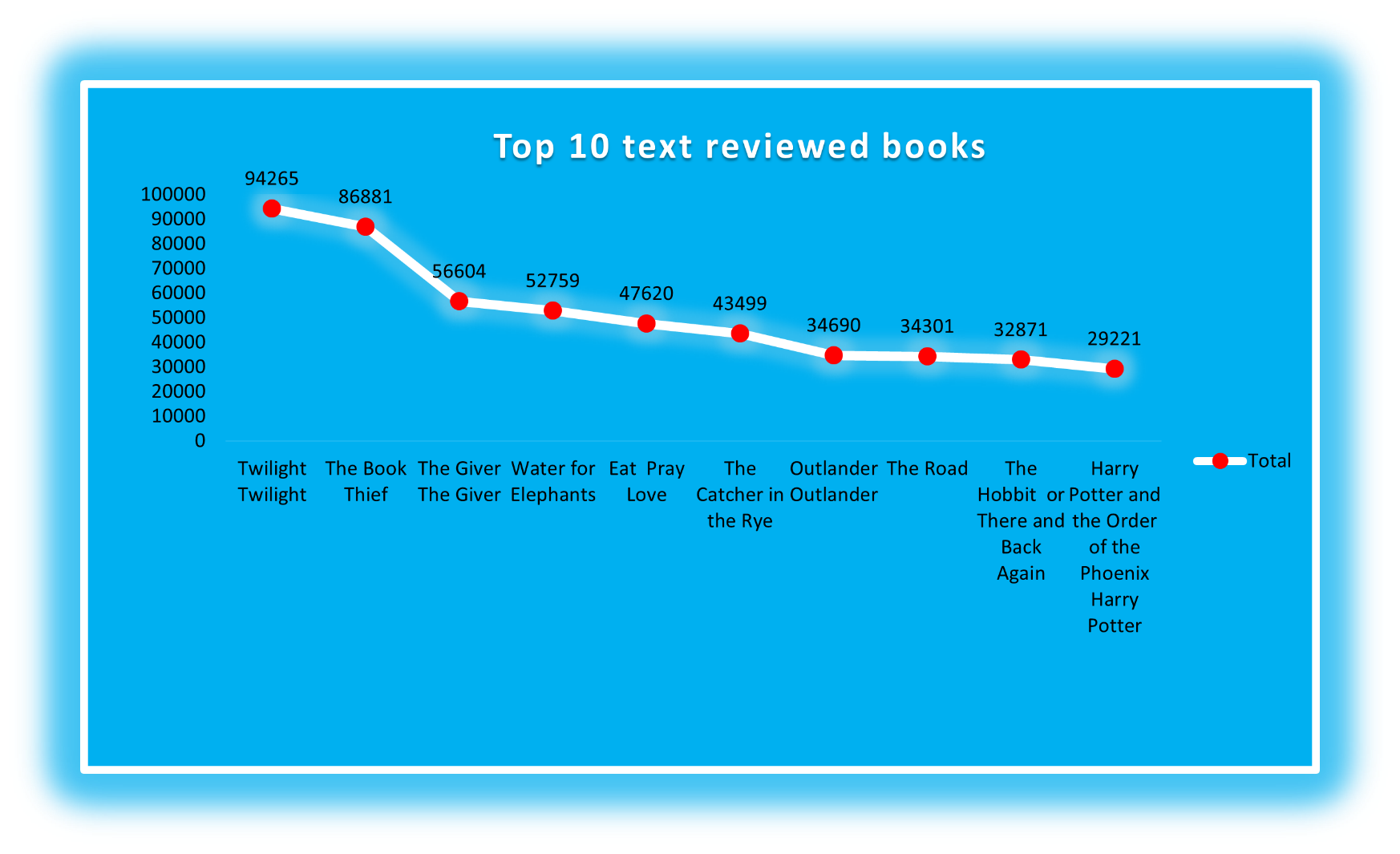
### Analysis results:

****

***Time line:***

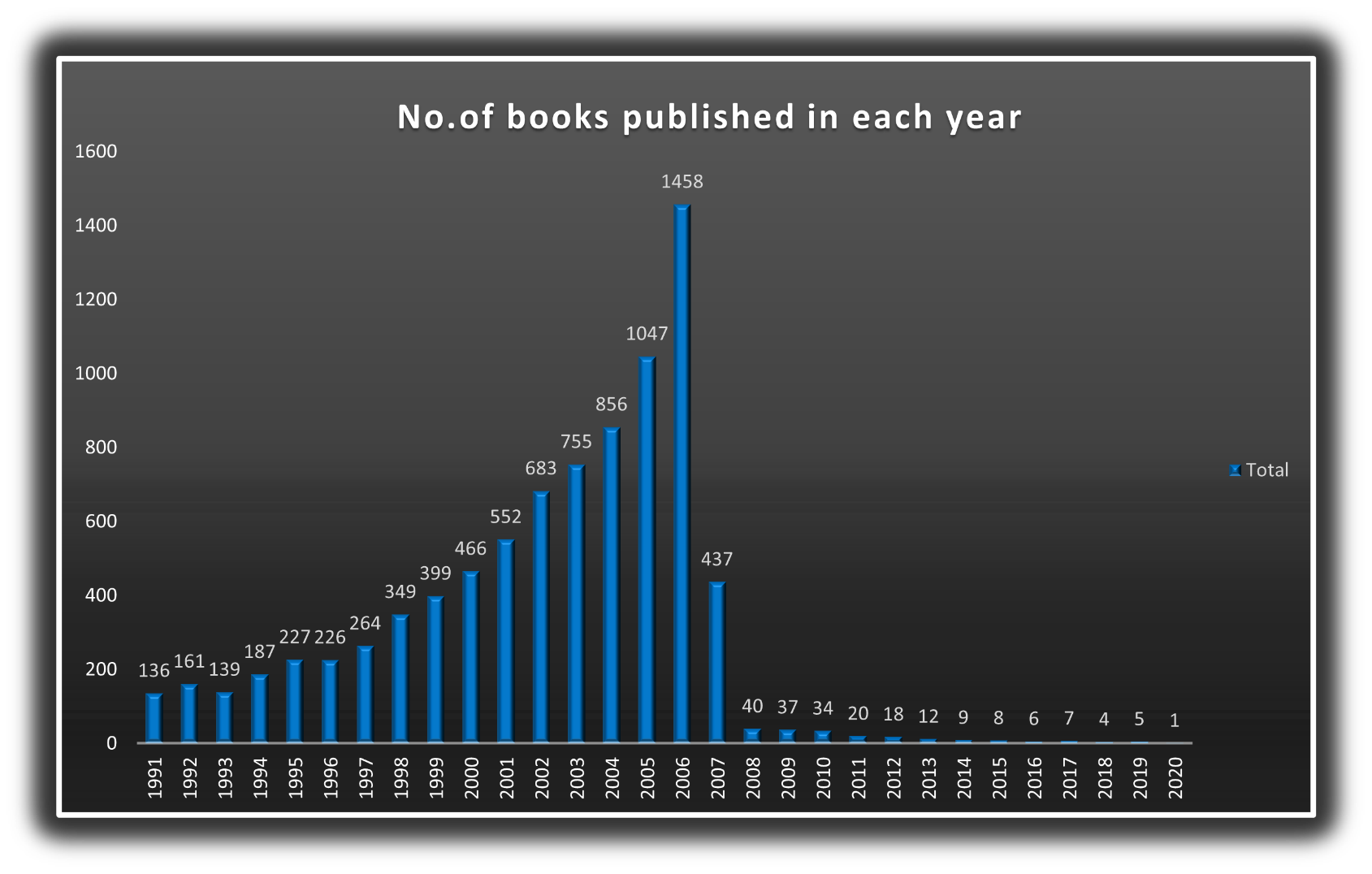
******

***Visualization:***

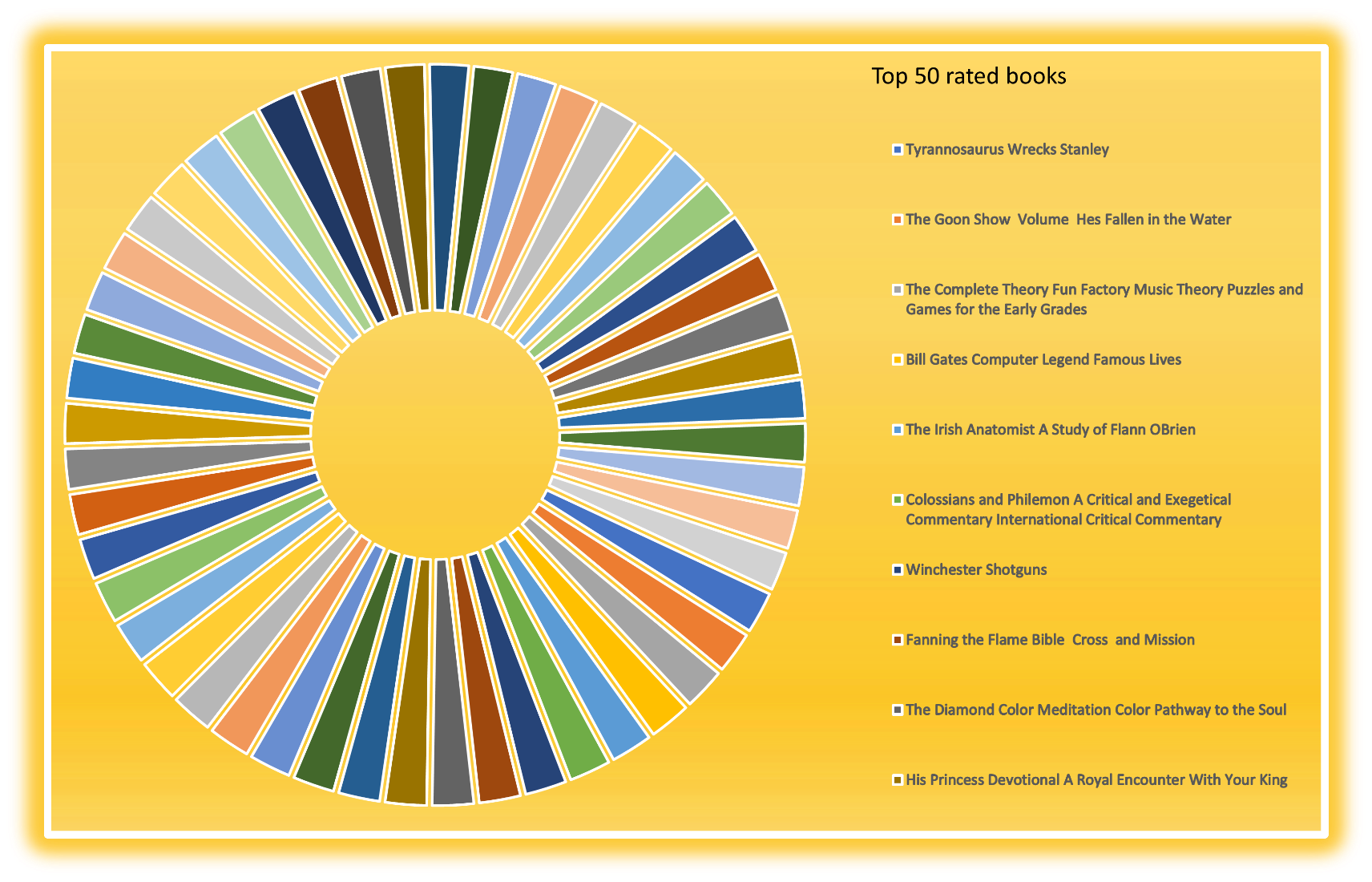
****

**List of Analysis with results**

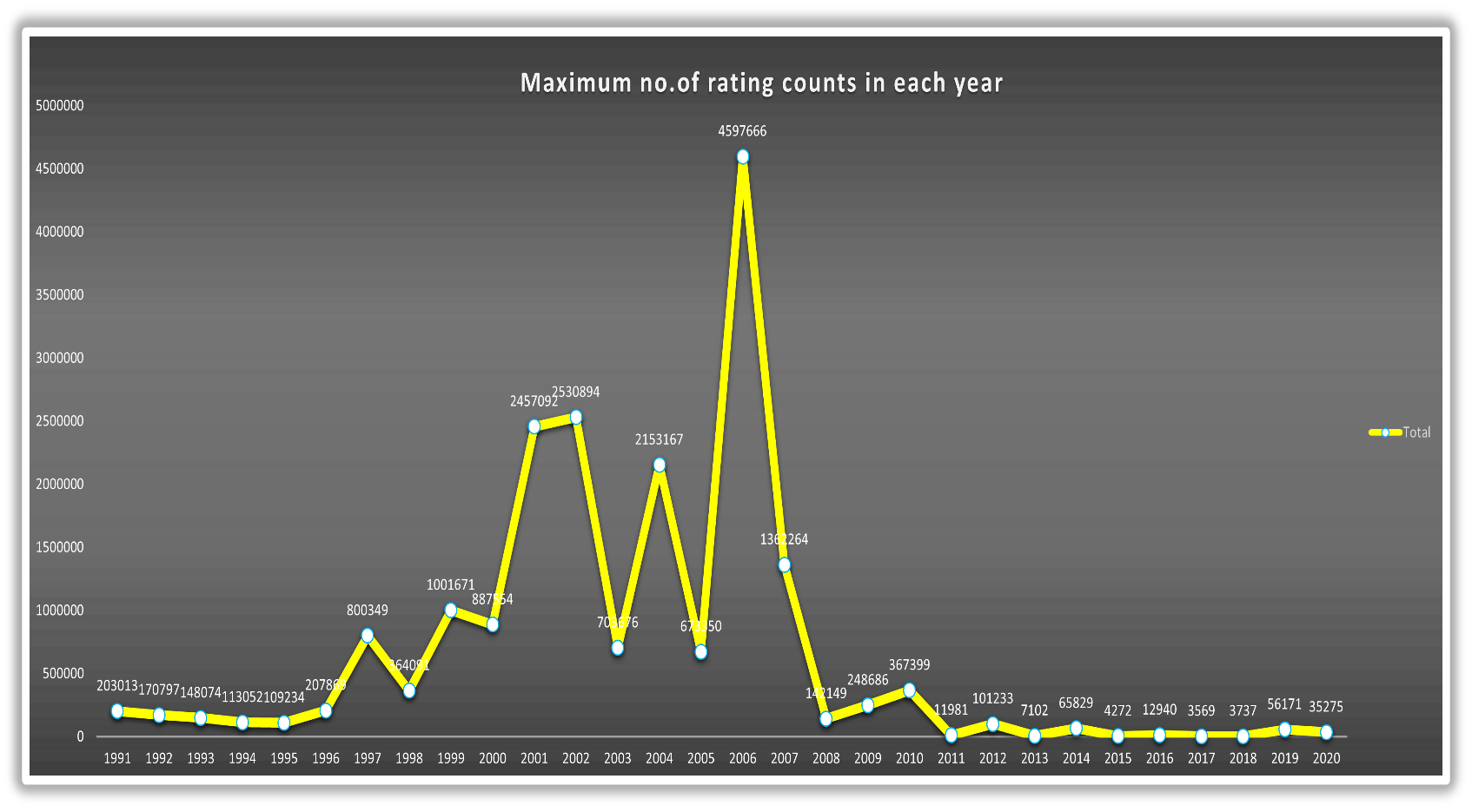
1. ***Total books published in each year:***



### Top 50 rated books:

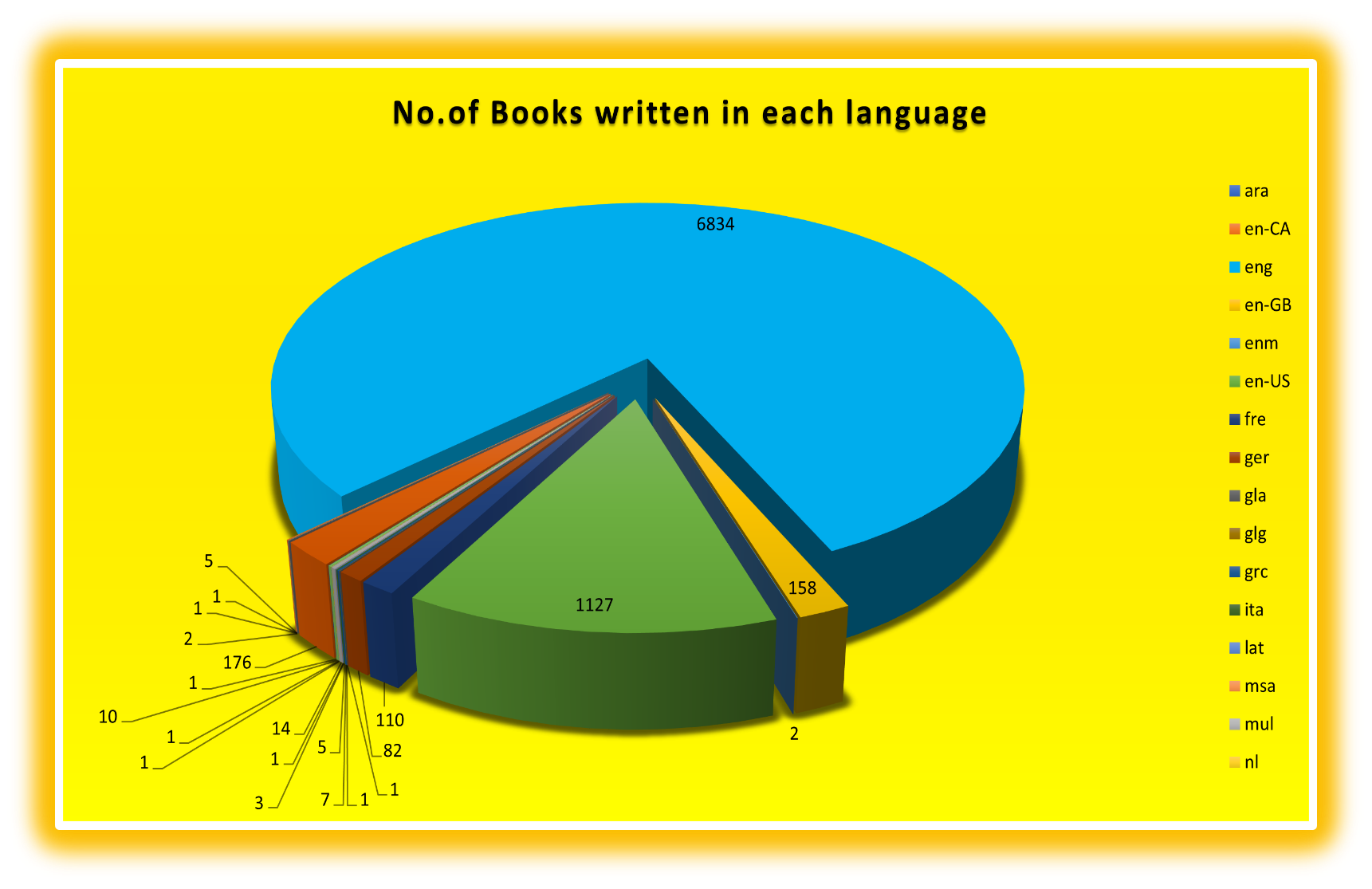


### Most rating counts containing books in each year:

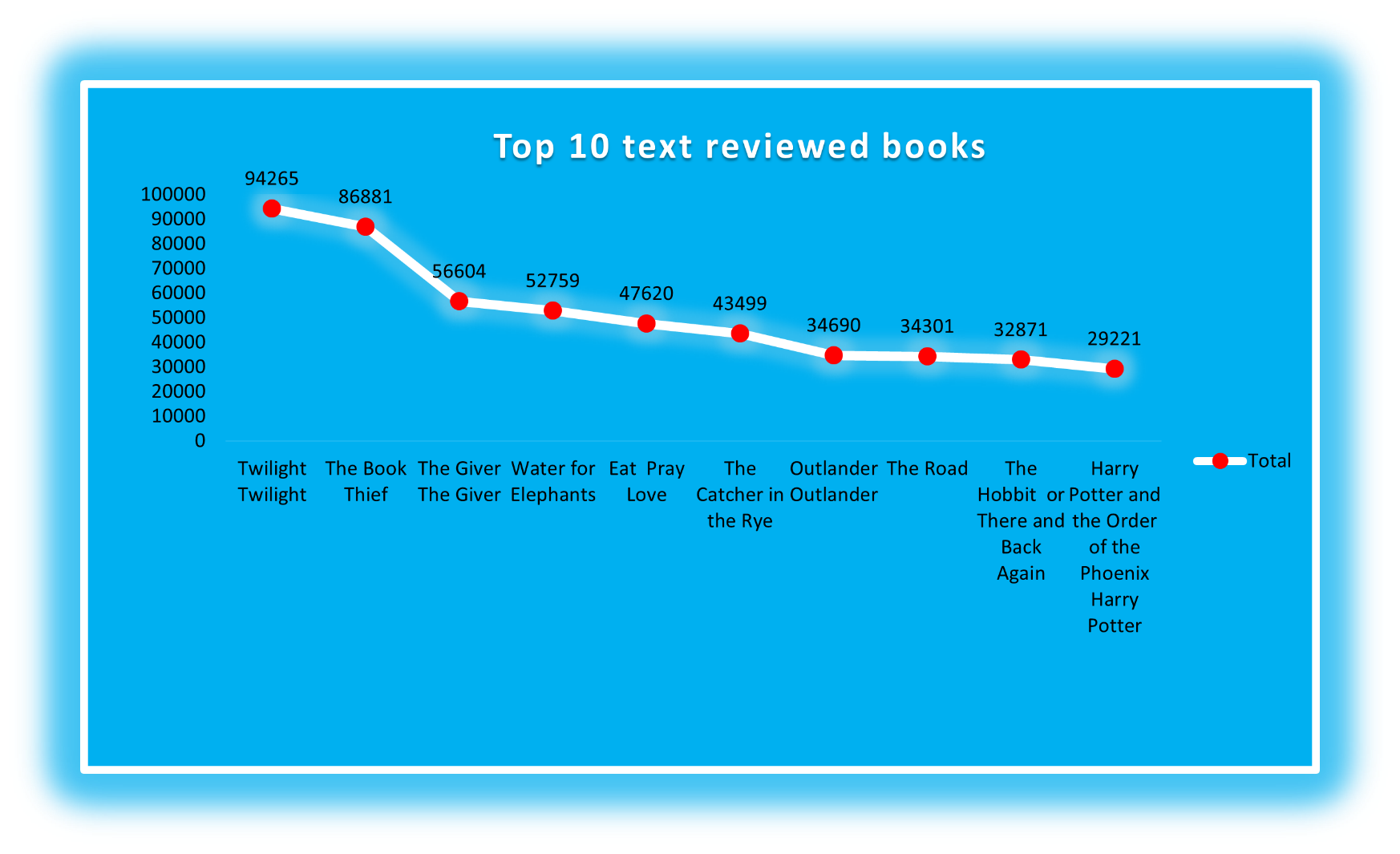


.

### No. of books written in each language:



### Top 10 text reviewed books:



# FINAL DASHBOARD:

**Graphical user interface

Description automatically generated**

**BIBLIOGRAPHY:**

* *Kaggle*
* *Trending videos of YouTube analysis*
* [*https://www.kaggle.com/datasnaek/youtube-new*](https://www.kaggle.com/datasnaek/youtube-new)
* [*https://www.analyticsvidhya.com/blog/2019/09/7-data-science- projects-github- showcase-your-*](https://www.analyticsvidhya.com/blog/2019/09/7-data-science-%20%20%20%20projects-github-%20%20%20%20showcase-your-) [*skills/*](https://www.analyticsvidhya.com/blog/2019/09/7-data-science-projects-github-showcase-your-skills/)
* [*https://en.wikipedia.org/wiki/YouTube*](https://en.wikipedia.org/wiki/YouTube)
* [*https://www.quora.com/How-does-Youtubes-algorithm-works-in-terms-of- trending-a-video*](https://www.quora.com/How-does-Youtubes-algorithm-works-in-terms-of-trending-a-video)